BioCreative VII Track 5 - LitCovid track Multi-label topic classification for COVID-19 literature annotation

Webinar

Qingyu Chen, Alexis Allot, Rezarta Islamaj,
Robert Leaman, Zhiyong Lu

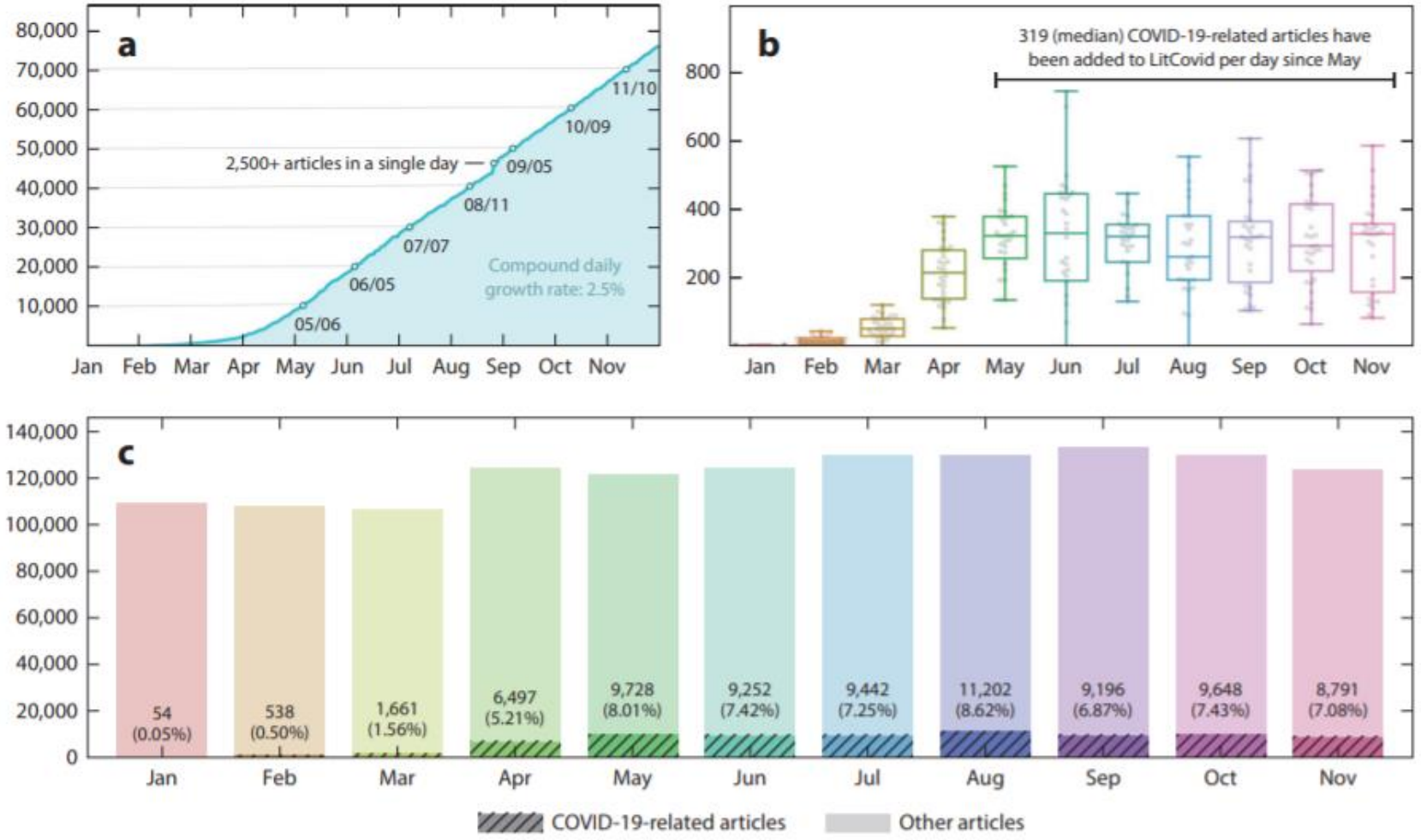
7/22/2021

Schedule

- Introduction
  - General background of LitCovid
  - The introduction to the BioCreative challenge
- Getting started
  - LitCovid Track
  - Timelines
  - Datasets and evaluations
- QA

Part 1 Introduction

COVID-19 literature growth



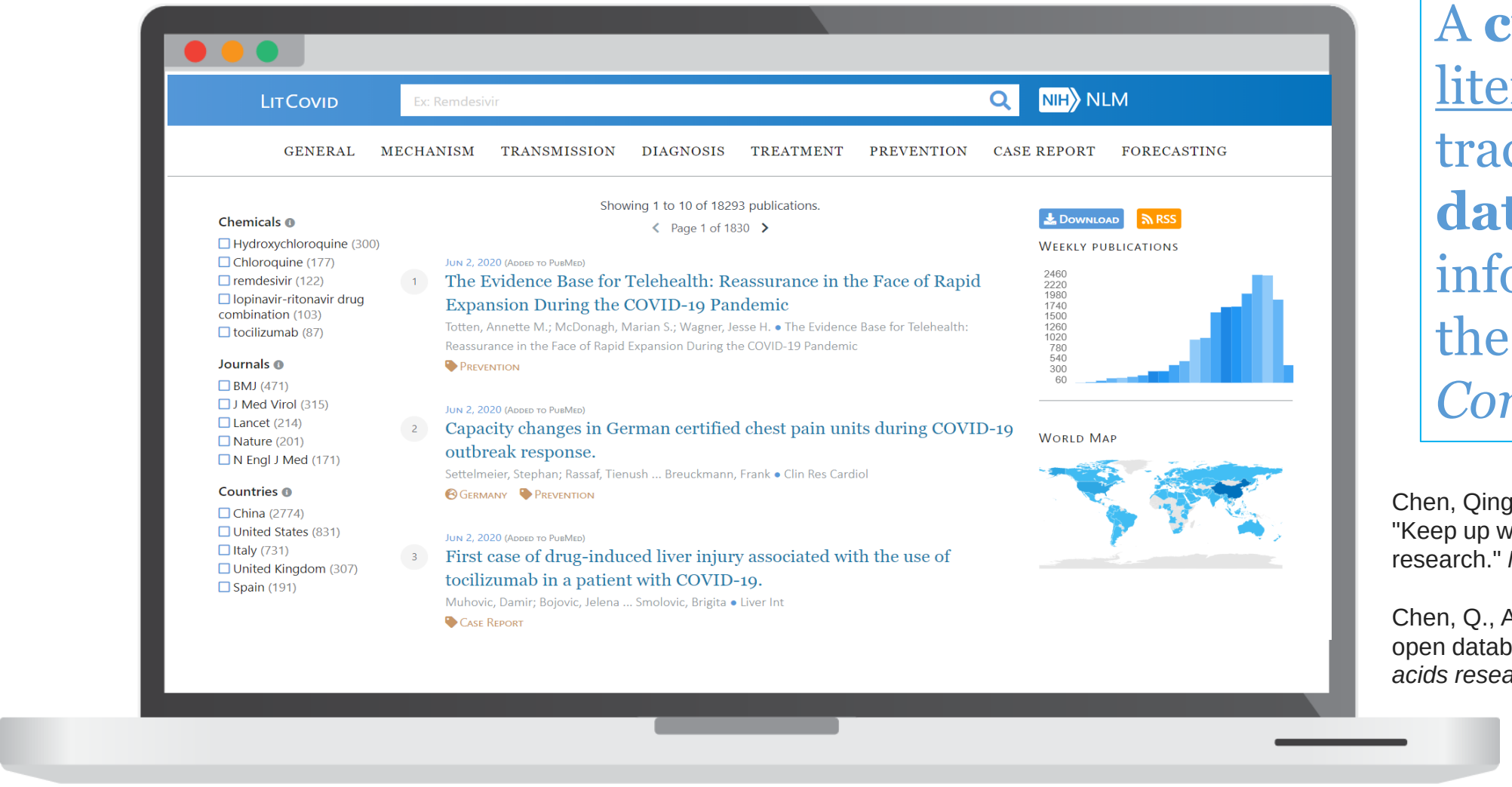
- ~10K COVID-19 related articles are deposited each month
- The peak was over 2.5K articles deposited in a single day
- Accounting for 7-8% of PubMed articles since May 2020

Chen, Qingyu, Robert Leaman, Alexis Allot, Ling Luo, Chih-Hsuan Wei, Shankai Yan, and Zhiyong Lu. "Artificial Intelligence in Action: Addressing the COVID-19 Pandemic with Natural Language Processing." Annual Review of Biomedical Data Science 4 (2021).

COVID-19: a taste of language variation

- pneumonia of unknown aetiology
- 2019-nCov infection
- novel coronavirus pneumonia
- SARS-CoV-2 infection
- COVID-19
- coronavirus disease 2019
- 2019 novel coronavirus infection disease
- nCOVID-19
- severe acute respiratory syndrome coronavirus 2 infection
- Coronavirus disease of 2019
- Wuhan coronavirus pneumonia
- CoV 19 infection
- COIVD-19 disease
- COVID-19 ARDS
- nCov-19 infection
- SARS-CoV-2 infectious disease
- CV-19
- coronavirus 2 syndrome
- SARS-CoV-2 associated ARDS
- C19

LitCovid: introduction

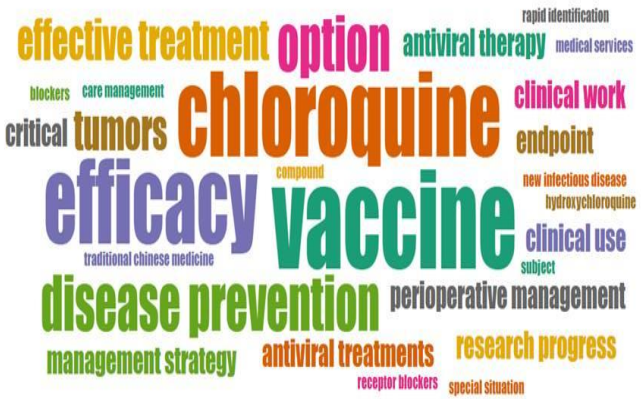


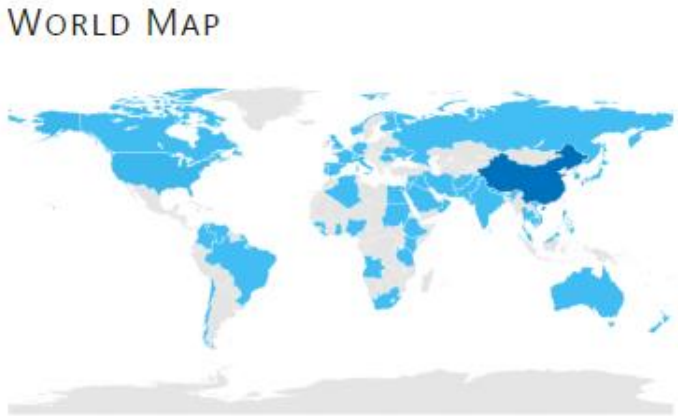
A **curated** literature hub for tracking **up-to-date** scientific information about the *2019 novel Coronavirus*

Chen, Qingyu, Alexis Allot, and Zhiyong Lu. "Keep up with the latest coronavirus research." *Nature* 579.7798 (2020): 193-193.

Chen, Q., Allot, A. and Lu, Z., 2021. LitCovid: an open database of COVID-19 literature. *Nucleic acids research*, *49*(D1), pp.D1534-D1540.

LitCovid: overview





1 Topics

Relevant publications are classified into main topics by a human annotator, such as : **General Info**, **Mechanism**, **Transmission**, **Treatment**, **Case Report**, **Epidemic Forecasting**

2 Weekly Overview

To better follow the **evolution of the research** on the epidemic, our users can view the **timeline** displaying the number of new publications per week.

3 Geographic Locations

We automatically extract country, region, city mentions from titles and abstracts of publications, allowing easy **search** and **filtering** of publications by **country**.

LitCovid: basic statistics

- Updates daily since Feb 2020
- Keeps track of over 150,000 COVID-19 related articles in PubMed
- Over 30 millions of hits by users worldwide for various information needs, such as evidence synthesis, drug discovery, and text and data mining
- Cross-referenced by hundreds of institutions in academia, government, and health organizations

Home dashboard

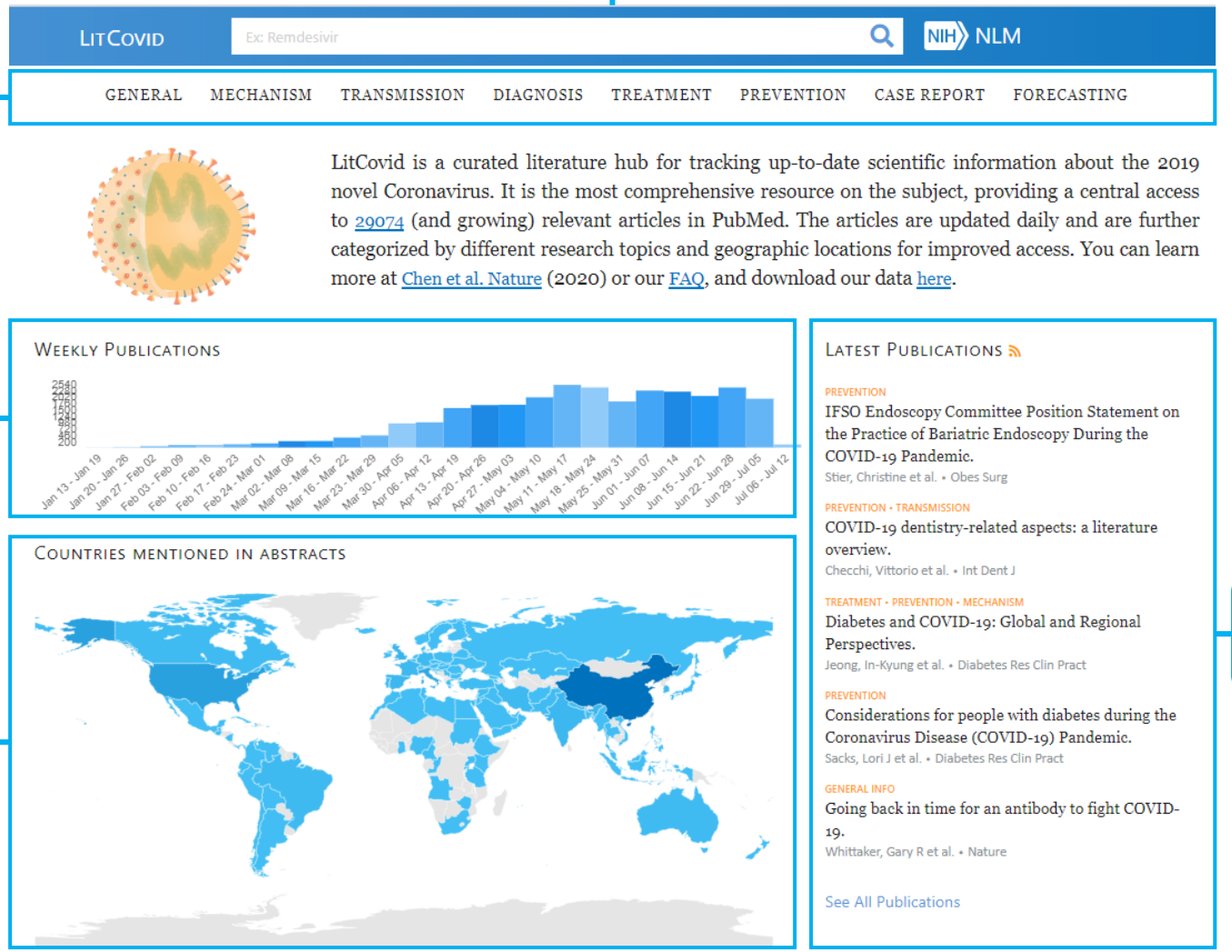
Search bar

Topics Navigation

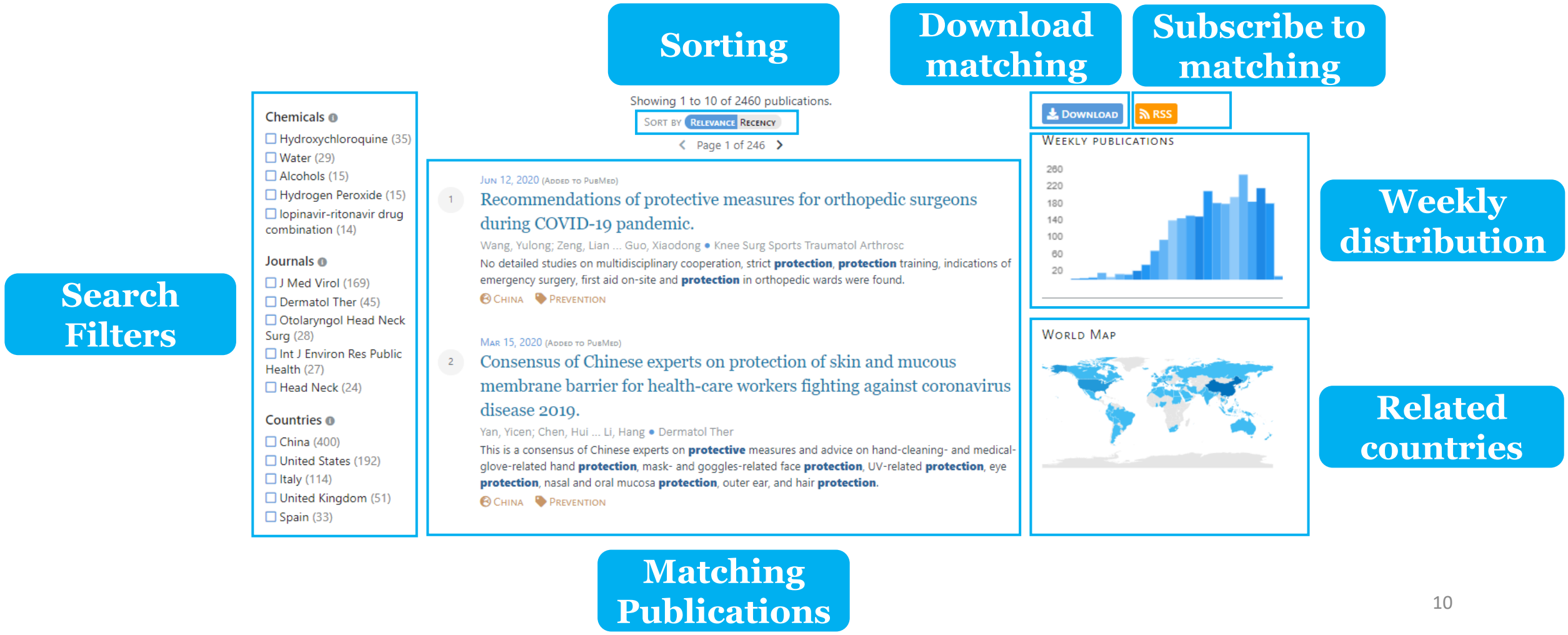
Weekly Publications

World Map

Latest Publications



Docsum Page

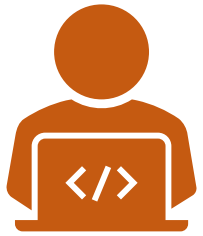

Sorting
Download matching
Subscribe to matching
Search Filters
Chemicals
Hydroxychloroquine (35)
Water (29)
Alcohols (15)
Hydrogen Peroxide (15)
lopinavir-ritonavir drug combination (14)
Journals
J Med Virol (169)
Dermatol Ther (45)
Otolaryngol Head Neck Surg (28)
Int J Environ Res Public Health (27)
Head Neck (24)
Countries
China (400)
United States (192)
Italy (114)
United Kingdom (51)
Spain (33)
Showing 1 to 10 of 2460 publications.
Sort by Relevance Recency
Page 1 of 246
Jun 12, 2020 (Added to PubMed)
1
Recommendations of protective measures for orthopedic surgeons during COVID-19 pandemic.
Wang, Yulong; Zeng, Lian ... Guo, Xiaodong • Knee Surg Sports Traumatol Arthrosc
No detailed studies on multidisciplinary cooperation, strict protection, protection training, indications of emergency surgery, first aid on-site and protection in orthopedic wards were found.
China Prevention
Mar 15, 2020 (Added to PubMed)
2
Consensus of Chinese experts on protection of skin and mucous membrane barrier for health-care workers fighting against coronavirus disease 2019.
Yan, Yicen; Chen, Hui ... Li, Hang • Dermatol Ther
This is a consensus of Chinese experts on protective measures and advice on hand-cleaning- and medical-glove-related hand protection, mask- and goggles-related face protection, UV-related protection, eye protection, nasal and oral mucosa protection, outer ear, and hair protection.
China Prevention
Download
RSS
Weekly Publications
260
220
180
140
100
60
20
Weekly distribution
World Map
Related countries
Matching Publications

Abstract Page

Download publications

1 Bibliography

Download LitCovid citations in **RIS** format for import into **reference management** software.



2 Scripts

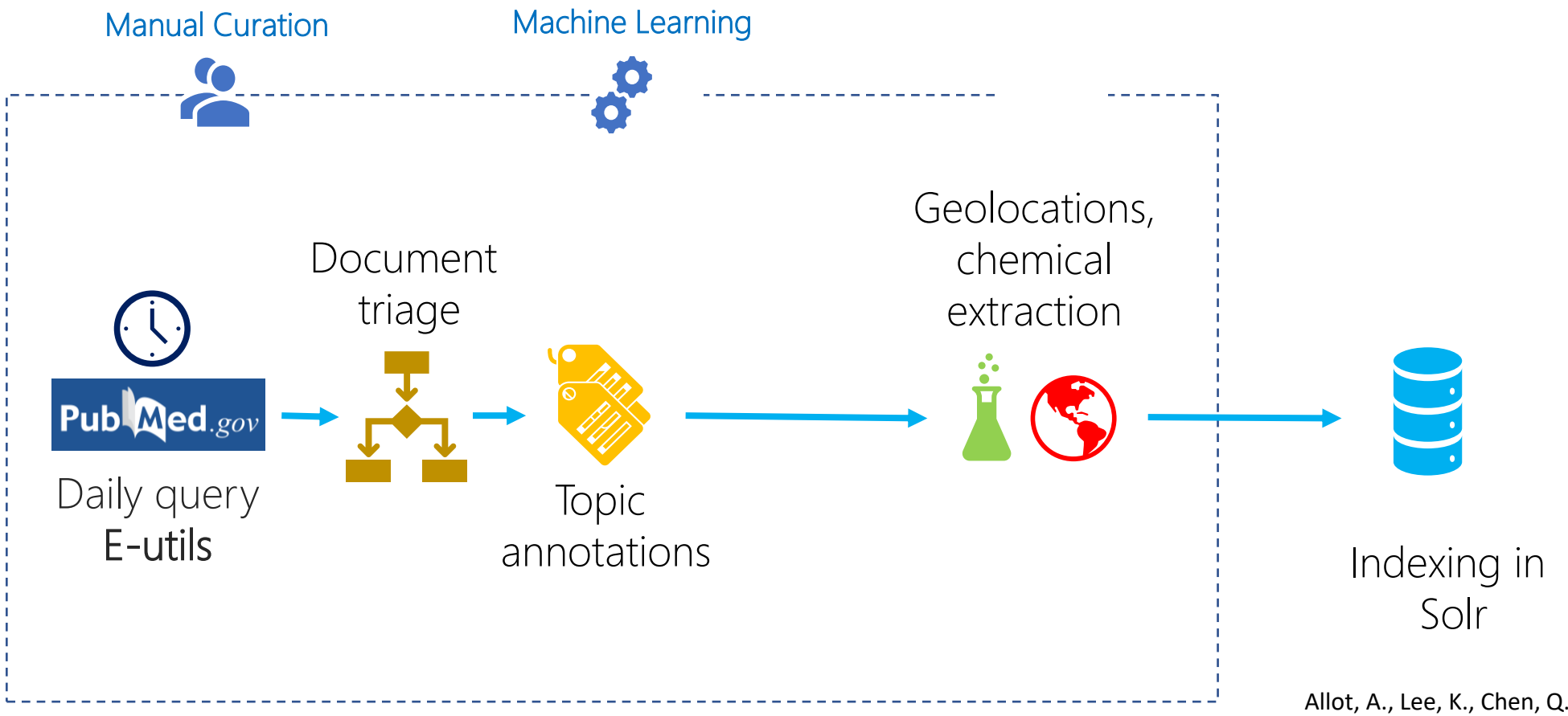
Download LitCovid citations in **TSV** format to process with **automated software**.

3 Text Mining

Download the text available from LitCovid articles, with **automatic annotations** by PubTator.

All publications or specific query

Curation pipeline



Allot, A., Lee, K., Chen, Q., Luo, L. and Lu, Z., 2021. LitSuggest: a web-based system for literature recommendation and curation using machine learning. Nucleic Acids Research.

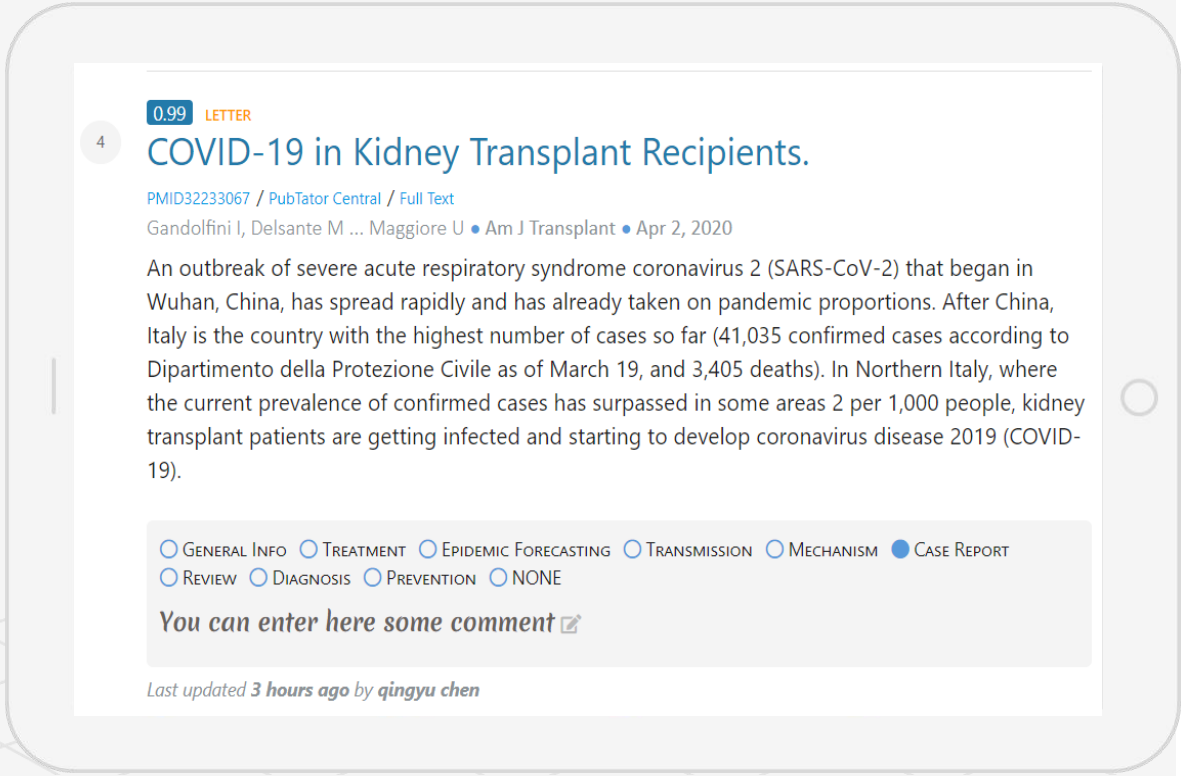
LitCovid daily curation pipeline

1. **Retrieve**
   A filtering query is automatically executed daily to retrieve new publications from PubMed.

2. **Classify**
   Machine-learning models are applied to classify whether a publication is related to COVID-19 and predict its topics if so.

3. **Curate**
   Human annotators manually verify the results from the last step.

4. **Export**
   Annotated publications are exported into a TSV file to be imported into LitCovid database.



Classifying publications

TWO STEPS PROCESS

1

A broad query allows to retrieve as much relevant publications as possible

("coronavirus"[All Fields] OR "ncov"[All Fields] OR "cov"[All Fields] OR "2019-nCoV"[All Fields] OR "COVID-19"[All Fields] OR "SARS-CoV-2"[All Fields]) AND ("2020/06/01"[CRDT] : "2020/06/01"[CRDT]) NOT preprint[pt]

2

ML & DL models are applied to classify whether they are related to COVID-19

We developed an ensemble of models, including: CNN, random forest and SVM.

PREPROCESSING

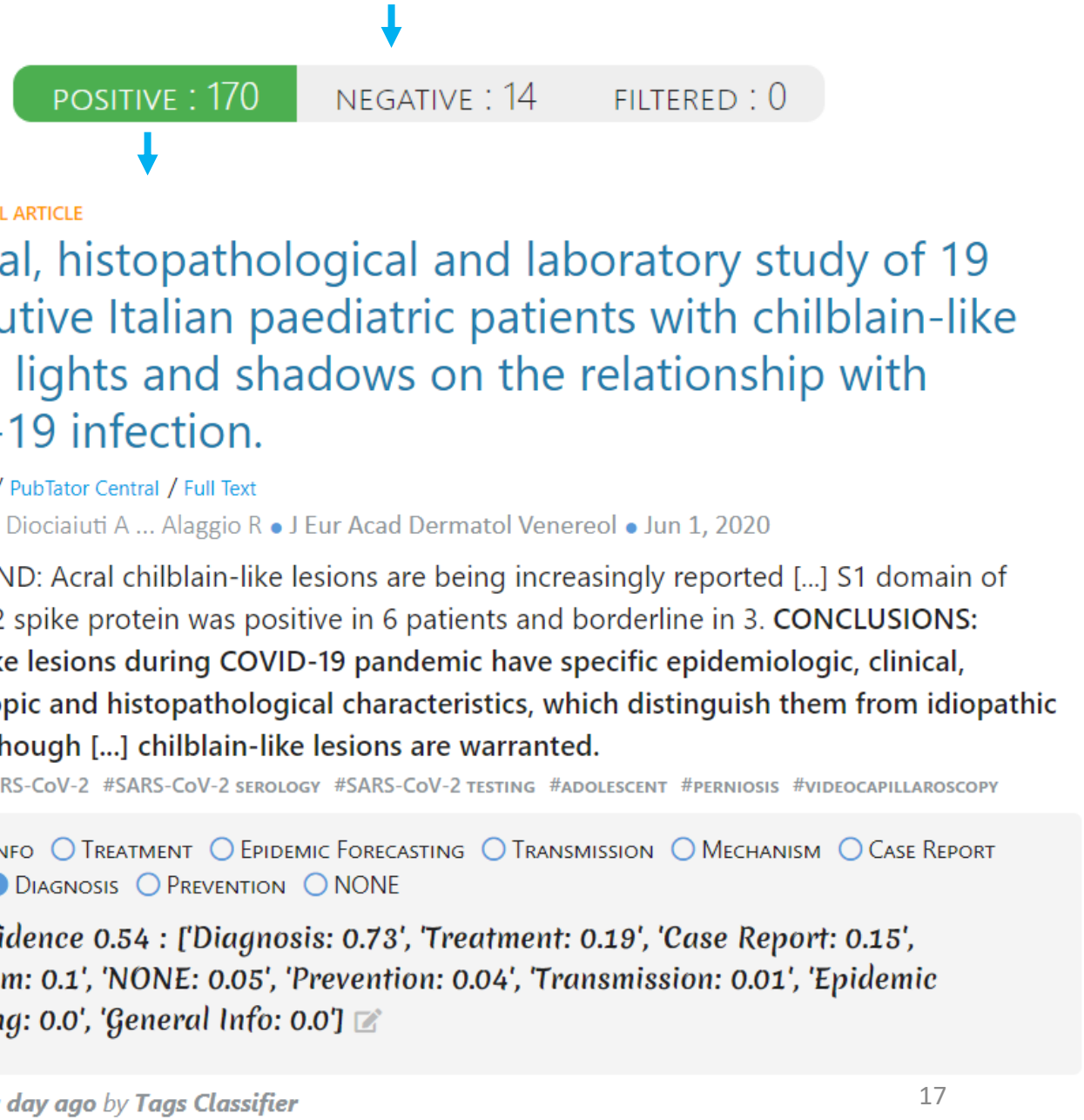
Classification evaluation

Precision: 0.99
Recall: 0.98
F-1: 0.98

Training/Test	6,360/1,591 articles

Assigning topics

- Binary BioBERT deep learning models were used



Topics evaluation

Precision: 0.78
Recall: 0.82
F-1: 0.80

Training/Test	17,845/4,461 articles

Geolocation tagging

- A pre-trained location tagging model from spacy is applied
- The predictions are post-processed using manually-crafted rules

Geolocation evaluation

Precision: 0.96
Recall: 0.93
F-1: 0.94

Training/Test N.A./312 articles

Automatic curation & manual curation in LitCovid

- LitCovid is updated daily, and this rapid growth significantly increases the burden of manual curation
- Currently, a combination of automatic and manual curation is used
- Topic annotation is the most challenging task in the curation pipeline which still requires manual curation efforts
- **Aim of the BioCreative LitCovid Track: call for a community effort to tackle automated topic annotation for COVID-19 literature**

BioCreative

www.biocreative.org

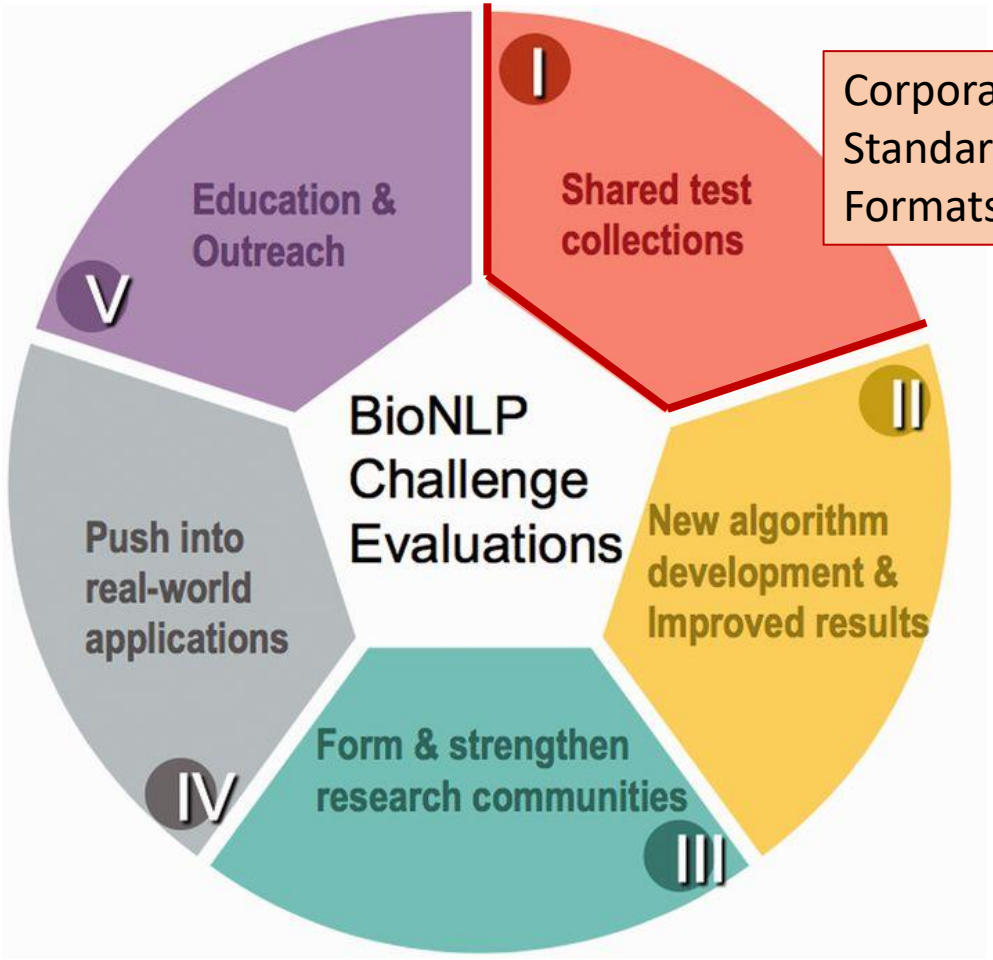
-Community-wide effort for evaluating text mining systems applied to the biomedical domain

-Collaborative and interdisciplinary effort

-Focuses on problems of importance to the biocuration, bioinformatics community and beyond

BioCreative VII challenge is underway

BioNLP Challenges impact/contributions



Corpora
Standards
Formats for NLP interoperability

Modified from Chung-Chi Huang, and Zhiyong Lu Brief Bioinform 2015;bib.bbv024

Part 2 Get started with the BioCreative LitCovid Track

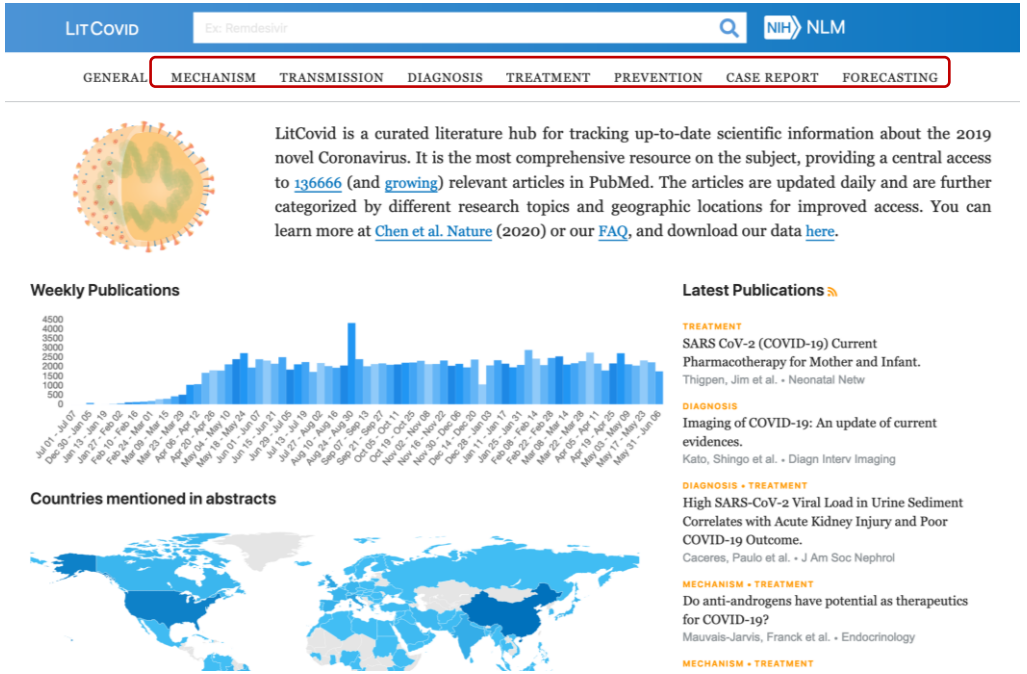
BioCreative LitCovid track

Need: Accelerate annotation of publication topics

LitCovid is an open database of COVID-19 literature

- **Task: Multilabel classification**

  Seven categories: mechanism, diagnosis, treatment, forecasting, case report, prevention, and transmission



https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-5/

Timeline

- **15th June:** Training and development set were released
- **15th July:** The evaluation script was released
- **22nd July:** Webinar
- **Late August:** Test set and submission instructions will be released
- **Early September:** Test set predictions due
- **Mid September:** Short technical systems description paper due
- **Late September:** Paper acceptance and review returned
- **TBD:** Database journal special issue

- The schedule might vary so please check the track page regularly. We will also send reminders and updates to the participants.

Datasets

Training and development datasets
Contain the publicly-available text of over 30 thousand COVID-19-related articles and their metadata (e.g., title, abstract, journal).

Categories are annotated with LitSuggest and reviewed by curators

Size: 25,088 PMIDs (training)
6,272 PMIDs (development)

Evaluation dataset

Contains articles that have been manually reviewed.

Size: 4,000 PMIDs (planned)

Association of hypercoagulation with severe acute respiratory syndrome coronavirus 2 infection.

PMID: 34083499 Jun 5, 2021

Nuthalapati, Poojith; Ghanta, Mohan Krishna; Natesh, Nagabhishek Sirpu; L V K S, Bhaskar •

Blood Res

Full Text DIAGNOSIS • MECHANISM • TREATMENT

The coronavirus disease 2019 (COVID-19) pandemic has emerged as a major threat to all healthcare systems across the globe, and it was declared a public health emergency of international concern by the World Health Organization (WHO). The novel coronavirus affects the respiratory system, producing symptoms such as fever, cough, dyspnea, and pneumonia. The association between COVID-19 and coagulation has been previously reported. Due to several inflammatory changes that occur in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections such as alterations in the levels of clotting factors, platelet activation leads to thrombus formation in coronary and cerebral vessels, leading to myocardial infarction and cerebrovascular accidents, respectively. Unfortunately, the progression of hypercoagulability in COVID-19 is rapid in patients with and without comorbidities. Hence, the proper monitoring of thrombotic complications in patients with COVID-19 is essential to avoid further complications. The implementation of guidelines for antithrombotic treatments based on the presentation of the disease is recommended. This review discusses the symptoms and mechanisms of upregulated coagulation in patients with COVID-19.

Keywords:
#covid-19 #d-dimer #hypercoagulation #sars-cov-2 #stroke #thrombosis

Chen et al., LitCovid: an open database of COVID-19 literature, doi: 10.1093/nar/gkaa952

Datasets

- The datasets can be accessed via https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/biocreative/

- It contains the following items:
  - Training dataset
  - Development dataset
  - A readme document:
    - Basic introduction
    - File format and descriptions of each field
    - **The label field contains annotated topics, i.e., the output**

- Let’s go through the datasets for demonstration

Datasets

- Each article can be assigned one or more labels (Treatment, Diagnosis, Prevention, Mechanism, Transmission, Epidemic Forecasting, and Case Report)

- Each label is separated by a semicolon, e.g., ‘Diagnosis;Treatment’ means that the article is assigned both the label Diagnosis and the label Treatment

- The test data will be provided in the same format, except the topic labels should be predicted by the participants

Datasets

- You can also use information beyond the prepared datasets as the additional inputs, for example:

  - MeSH terms
  - Biomedical entities
  - Cross-referenced annotations

Evaluation scripts

- Evaluation scripts can be accessed via https://github.com/ncbi/biocreative_litcovid

- It contains the following items:
  - A detailed step-by-step instruction to set up
  - Evaluation scripts
  - Sample documents
    - prediction_label_samples.csv: you will provide the test prediction file in exact format

- Let’s go through the files for demonstration

Summary

- Getting started:
  - Download the datasets and explore the data
  - Download the evaluation and verify whether you can reproduce the results
  - Check the sample prediction file format

- Resources:
  - The track page (https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-5/) contains all the information about the track
  - The BioCreative page (https://biocreative.bioinformatics.udel.edu/) contains the overall information about the challenge tasks
  - Registration: https://docs.google.com/forms/d/e/1FAIpQLScdMnKFMncL8qDkcRx6aV6lYRm8PbufPs1rIAODwxCcPoLkcg/viewform

Summary

- Questions:
  - Please contact qingyu.chen@nih.gov with the subject heading "BioCreative Track 5 LitCovid questions”
  - We will provide the answers to common questions to the FAQ section of the track page
  - We will also send updates to the participants regularly
- The slides and recording will also be available at the dataset folder
- On behalf of the organizers, we thank you for attending the webinar and look forward to your participations!

Questions?

Acknowledgment

- This research is supported by the NIH Intramural Research Program, National Library of Medicine

- Some slides are adapted from Zhiyong Lu, Alexis Allot, and Cecilia Arighi